POPE

Negative Sampling

Nonexistent objects

***Random**:* dog、 apple, ...

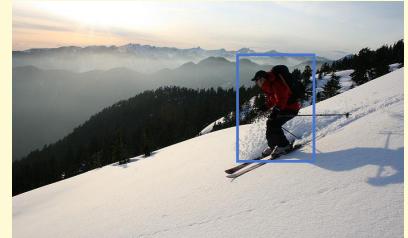***Popular**:* table, knife, ...

***Adversarial**:* snowboard...

Q₁: Is there a skis?
A₁: (GT = yes)
Q₂: Is there a person?
A₂: (GT = yes)
Q₃: Is there a backpack?
A₃: (GT = yes)

Q₁: Is there a dog?
A₁: (GT = no)
Q₂: Is there a table?
A₂: (GT = no)
Q₃: Is there a snowboard?
A₃: (GT = no)

Only Text-axis evaluation

BEAF

Object Removal

Q₁: Is there a skis?
A₁: (GT = yes)
Q₂: Is there a person?
A₂: (GT = yes)
Q₃: Is there a snowboard?
A₃: (GT = no)

Q₁: Is there a skis?
A₁: (GT = yes)
Q₂: Is there a person?
A₂: (GT = no)
Q₃: Is there a snowboard?
A₃: (GT = no)

Text & Image-axis Evaluation

Causal-HalBench
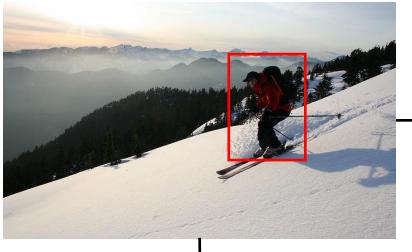
Original image
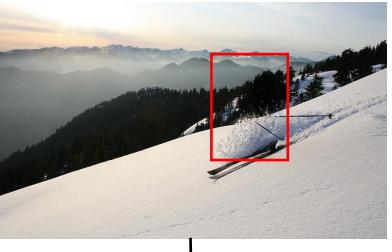

Counterfactual inpainting

Counterfactual image

Q₁: Is there a skis?
A₁: (GT = yes)
Q₂: Is there a person?
A₂: (GT = yes)
Q₃: Is there a snowboard?
A₃: (GT = no)

Q₁: Is there a skis?
A₁: (GT = yes)
Q₂: Is there a bear?
A₂: (GT = yes)
Q₃: Is there a snowboard?
A₃: (GT = no)

Q₁: Is there a skis?(Asking about unchanged objects in the image, i.e., contextual objects.)

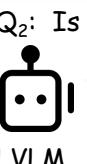LVLM → A₁:yes ✅ → A₁:no ❌ → **CAC:** Contextual object Accuracy Change

Q₂: Is there a bear?(Asking about inpainted objects , i.e., counterfactual objects.)

LVLM → A₂:no ❌ → **CHR:** Counterfactual object Hallucination Rate

Q₃: Is there a snowboard?(Asking about absent objects)

LVLM → A₃:yes ❌ → A₃:no ✅ → **AAC:** Absent object Accuracy Change

Causal-Powered Text & Image-axis Evaluation